Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 3419–3423, April 1992
Genetics

Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences

(marker-selected libraries/CA repeats/sequence-tagged sites/genetic mapping/dog genome)

ELAINE A. OSTRANDER*†‡, PAM M. JONG*†, JASPER RINE*†, AND GEOFFREY DUYK‡§

*Department of Molecular and Cellular Biology, 401 Barker Hall, University of California, Berkeley, CA 94720; †Human Genome Center, Lawrence Berkeley Laboratory, 1 Cyclotron Road, 74-157, Berkeley, CA 94720; and §Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115

Communicated by Philip Leder, January 8, 1992

ABSTRACT **We describe an efficient method for the construction of small-insert genomic libraries enriched for highly polymorphic, simple sequence repeats. With this approach, libraries in which 40–50% of the members contain $(CA)_n$ repeats are produced, representing an ≈50-fold enrichment over conventional small-insert genomic DNA libraries. Briefly, a genomic library with an average insert size of less than 500 base pairs was constructed in a phagemid vector. Amplification of this library in a *dut ung* strain of *Escherichia coli* allowed the recovery of the library as closed circular single-stranded DNA with uracil frequently incorporated in place of thymine. This DNA was used as a template for second-strand DNA synthesis, primed with $(CA)_n$ or $(TG)_n$ oligonucleotides, at elevated temperatures by a thermostable DNA polymerase. Transformation of this mixture into wild-type *E. coli* strains resulted in the recovery of primer-extended products as a consequence of the strong genetic selection against single-stranded uracil-containing DNA molecules. In this manner, a library highly enriched for the targeted microsatellite-containing clones was recovered. This approach is widely applicable and can be used to generate marker-selected libraries bearing any simple sequence repeat from cDNAs, whole genomes, single chromosomes, or more restricted chromosomal regions of interest.**

The use of DNA sequence polymorphisms as codominant markers for the construction of genetic maps has enhanced the feasibility of genetic analysis of organisms with large, complex genomes (1). In the absence of biochemical characterization of a gene product, the use of the genetic map position for a human disease gene or other complex phenotypes has been the primary route for isolating the gene. However, mammalian genetics in general and human genetics in particular are limited by the availability of informative genetic markers. An ideal set of markers would be homogeneously distributed, highly informative, easily utilized, and readily transferred between laboratories. Microsatellite repeat sequences, such as $(CA)_n$ or $(TG)_n$, satisfy these criteria (2, 3). This class of repeats is very abundant, occurring at a frequency of once every 50–150 kilobases (kb) in mammalian genomes (4–7). Individual loci frequently have multiple alleles in which the degree of length heterogeneity appears to correlate with the length of the repeat (8). In the case of $(CA)_n$ repeats, loci with 16 or more repeat units typically have polymorphic information content (PIC) values of 0.5 or greater. These length polymorphisms are detected by PCR-based assays that define a polymorphic sequence-tagged site (STS) (9). STS-based markers are useful for joining genetic maps to physical maps and are currently being used as the foundation for genetic maps of whole organisms, as well as for mapping specific loci.

Generation of a high-density map of markers for an entire genome or a single chromosome requires the isolation and characterization of hundreds of markers such as microsatellite repeats (10, 11). Two simple yet tedious approaches have generally been used for this task. One approach is to screen a large-insert genomic library with an end-labeled $(CA)_n$ or $(TG)_n$ oligonucleotide ($n > 15$). Clones that hybridize to the probe are purified and divided into subclones, which are screened by hybridization for a fragment containing the repeat. The fragment is then sequenced, and a STS is created by choosing unique primers that flank the repeat and produce a fragment of convenient, discrete size upon amplification by PCR. The drawbacks of this approach are the requirement for many blot hybridizations and the difficulty of sequencing the relatively large subclones. An alternative approach is to construct and screen a small-insert [200–500 base pairs (bp)] genomic library constructed in a plasmid vector. The expected frequency of $(CA)_n$ repeats in this small-insert library is low, about 1 per 100–400 colonies. Consequently, large numbers of plates must be screened at relatively low densities to obtain a significant pool of markers.

To overcome the limitations of these approaches, we developed an efficient method for genetic selection of small-insert genomic DNA libraries that are highly enriched for microsatellite sequences. In these libraries, nearly 50% of the members contain long microsatellite repeats, allowing for the recovery of hundreds of potential genetic markers by screening the colonies on a single Petri plate.

MATERIALS AND METHODS

Construction of Genomic Libraries. The primary library was constructed with canine genomic DNA purified from dog spleen by standard protocols (12). Genomic DNA was digested with a mixture of restriction enzymes including *Sau*3A1, *Rsa* I, *Hae* III, *Eco*RV, and *Ssp* I. The ends of the fragments were repaired with the Klenow fragment of *Escherichia coli* DNA polymerase I and then ligated into the *Sma* I site of pBluescript KS(+) (Stratagene). The ligation mixture was extracted with phenol/chloroform/isoamyl alcohol (25:24:1), the nucleic acids were precipitated with 0.3 M sodium acetate and 2 volumes of ethanol, and the resuspended material was digested with *Sma* I to linearize any pBluescript vector that lacked insert. Portions of this mixture were transformed by electroporation into *E. coli* XL1-Blue {*recA1, endA1, gyrA96, thi, hsdR17, supE44, relA1, lac* [F′ *proAB, lacI*^{q}*ZΔM15*, Tn*10* (Tet^R)]} and BSJ72 [*recA*$^+$, *hsdΔ5*, λ^R, Str^R, r^-m^-, Δ*lac–pro* (F′ *traD36, proAB, lacI*^{q}*ZΔM15*)] cells (Bio-Rad Gene Pulser). "Miniprep" DNA was prepared from 10 random colonies by an alkaline

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: STS, sequence-tagged site; PIC, polymorphic information content.

‡To whom reprint requests should be addressed.

lysis procedure (12). Analysis of miniprep DNA cleaved by *Pvu* II demonstrated that each clone contained an insert.

Construction of Marker-Selected Libraries. Microsatellite libraries were constructed by first transforming portions of the ligation mixture described above into CJ236 cells [F⁻ *dut-1 ung-1 thi-1 relA1*/pCJ105 (Cm^R)] (13, 14). Five thousand to 7500 colonies were plated on each of eight LB plates (150 × 15 mm) containing ampicillin (100 μg/ml; LB-Amp). After an overnight incubation, 10 ml of 2× YT plus ampicillin (100 μg/ml) was added to the surface of each plate. The plates were rotated gently at 37°C in a Lab-Line Instruments Environ Orbit-Shaker for 2 hr, after which the resulting cell suspensions were collected. The plates were rinsed with a second aliquot of 2× YT, the aliquots were combined, and the total mixture was washed once with 10 ml of 2× YT. Ten ml of LB-Amp was then inoculated with 50 μl of washed colony mix and grown to saturation at 37°C. Aliquots (10 μl) of this culture were used to inoculate each of twenty-four 2-ml cultures in 2× YT. Single-stranded DNA was prepared from each of these cultures by using the procedure described by Vieira and Messing (15), with slight modifications.

Primer extension reactions were carried out by combining 3 μg of single-stranded DNA, 20 pmol of the specified $(CA)_n$ or $(TG)_n$ primer, 10 μl of 10× PCR buffer (Perkin–Elmer/Cetus), and 200 μM each dNTP in a final volume of 100 μl. Samples were heated at 94°C for 10 min and then cooled to 75°C for extension reactions using 5′ phosphorylated $(CA)_n$ or $(TG)_n$ primers in which $n = 10$ or 15, or to 78°C when $n = 20$. Five units of *Taq* DNA polymerase (Perkin–Elmer/Cetus) was then added to each sample and the mixture was incubated for an additional 30 min at the temperature indicated. The reactions were terminated by extraction with 1 volume of phenol/chloroform/isoamyl alcohol (25:24:1), followed by extraction with chloroform/isoamyl alcohol (24:1), and ethanol precipitation with 20 μg of glycogen, 0.3 M sodium acetate, and 3 volumes of ethanol. Precipitates were resuspended in 10 mM Tris, pH 8.0/1 mM EDTA and incubated with T4 DNA ligase (New England Biolabs) at 37°C in the buffer provided by the supplier, in a final volume of 50 μl. One microliter of the ligation mixture, equivalent to 0.06 μg of the single-stranded template DNA, was transformed by electroporation into either XL1-Blue or BSJ72 cells, yielding ≈3800 colonies (≈6.3×10^4 colonies per microgram of input single-stranded template DNA). These transformants were referred to as the marker-selected library.

Screening of Primary and Marker-Selected Libraries for $(CA)_n$ Repeat-Containing Clones. To determine the frequency of $(CA)_n$-containing clones in each library, both the primary and marker-selected libraries were initially transformed by electroporation into XL1-Blue or BSJ72 cells and plated on LB-Amp plates. Individual colonies were picked at random and patched onto fresh LB-Amp plates. Colonies were grown for 3 hr at 37°C and then lifted onto nitrocellulose filters (Schleicher & Schuell). Filters were placed colony-side-up on a fresh set of LB-Amp plates and incubated at 37°C until individual patches were confluent. Filters were screened with a $(CA)_{15}$ oligonucleotide as described in Fig. 1.

PCR. Primers for PCRs were selected by using the program Primer 0.5, kindly provided by Stephen Lincoln, Mark Daly, and Eric Lander (Whitehead Institute, Cambridge, MA; unpublished). Primers were 18–22 bp long and were chosen to give PCR products in the range 100–225 bp. Each PCR mixture contained either 1 μg of genomic DNA or 0.25 μg of plasmid DNA, 20–40 pmol of unlabeled primer, 20–40 pmol of primer previously end-labeled with [γ-^{32}P]ATP, 200 μM dNTPs, and 3 μl of 10× PCR buffer (Perkin–Elmer/Cetus), in a total volume of 30 μl. Samples were overlaid with 50 μl of mineral oil, heated to 94°C for 10 min, and cooled to 75°C, and 0.8 unit of *Taq* DNA polymerase was added. Samples were immediately amplified in an MJ thermocycler (MJ Research, Cambridge, MA) for 26 cycles consisting of denaturation at 94°C (1 min), annealing at 60°C (1 min), and extension at 74°C (1 min). The final extension step was for 5 min, and the samples were cooled and stored at −20°C. Three microliters of reaction product was mixed with 3–4 μl of sequence loading buffer, and the samples were boiled for 5 min, cooled on ice, and loaded onto a 6% polyacrylamide sequencing gel for electrophoresis. Gels were autoradiographed, without drying, at −80°C with an intensifying screen for 12–18 hr.

RESULTS

Construction of Marker-Selected Libraries Highly Enriched for $(CA)_n$ Repeats. Marker-selected libraries highly enriched for clones that contain $(CA)_n$ repeats were constructed from a dog genomic library as follows. The primary genomic library was constructed in phagemid vectors and then propagated in a bacterial strain deficient in dUTPase (*dut* gene product) and uracil-N-glycosylase (*ung* gene product). In the absence of dUTPase, dUTP can compete effectively with dTTP for incorporation into DNA. When a *dut* mutation is combined with a mutation in uracil-N-glycosylase, an enzyme which normally removes deoxyuridine, DNA containing high levels of uracil are allowed to accumulate. Following superinfection of the primary library in a *dut ung* strain with M13 helper phage, circular single-stranded DNA containing uracil was isolated. The circular single-stranded DNA molecules were converted to circular double-stranded DNA by *in vitro* primer extension using $(CA)_n$ or $(TG)_n$ oligonucleotides and *Taq* DNA polymerase. In these experiments the number of repeats (n) in the primer was 10, 15, or 20 as specified. The products of this primer extension reaction were transformed into an *E. coli* strain with wild-type alleles at the *dut* and *ung* loci. This provides a strong genetic selection favoring the replication of the primer-extended products for two reasons: (*i*) circular single-stranded DNA transforms with significantly lower efficiency than circular double-stranded DNA (13, 14); (*ii*) uracil-containing DNA is degraded because the uracil-N-glycosylase contained in the wild-type strain removes the incorporated uracil residue, leaving an unstable sugar-phosphate backbone that is susceptible to scission by specific nucleases, creating a block to DNA replication (16). The circular double-stranded DNA products are rescued because the thymidine-containing primer-extended strand serves as a template for repair synthesis following the excision of uracil from the complementary strand. The resultant library is highly enriched for microsatellite containing clones that contain the targeted $(CA)_n$ repeats.

Screening of Primary and Marker-Selected Libraries for $(CA)_n$ Repeats. Random clones from both the primary and marker-selected libraries were screened for the presence of $(CA)_n$-containing isolates (Fig. 1). Each of the three filters in the right-hand column contained 100 randomly chosen clones from the primary dog library. A total of 3 positive colonies were detected. Each of the three filters in the left-hand column contained 100 randomly chosen colonies from a marker-selected library in which primer extension reactions were done with a $(CA)_{15}$ oligonucleotide. In each of the filters, 40–50% of the clones were strongly positive. Marker-selected libraries constructed by using a $(CA)_{20}$ primer gave the same result. These data suggest that constructing small-insert genomic libraries by a marker-selection procedure enriches for $(CA)_n$ repeats by ≈50-fold.

Average Length of $(CA)_n$ Repeats in Primary and Marker-Selected Libraries. The best indicator of informativeness for a $(CA)_n$ microsatellite is its longest run of uninterrupted repeats (8). Fifteen to 20 positive colonies were randomly selected from each screen and characterized by DNA sequence analysis to determine the average repeat length (Table

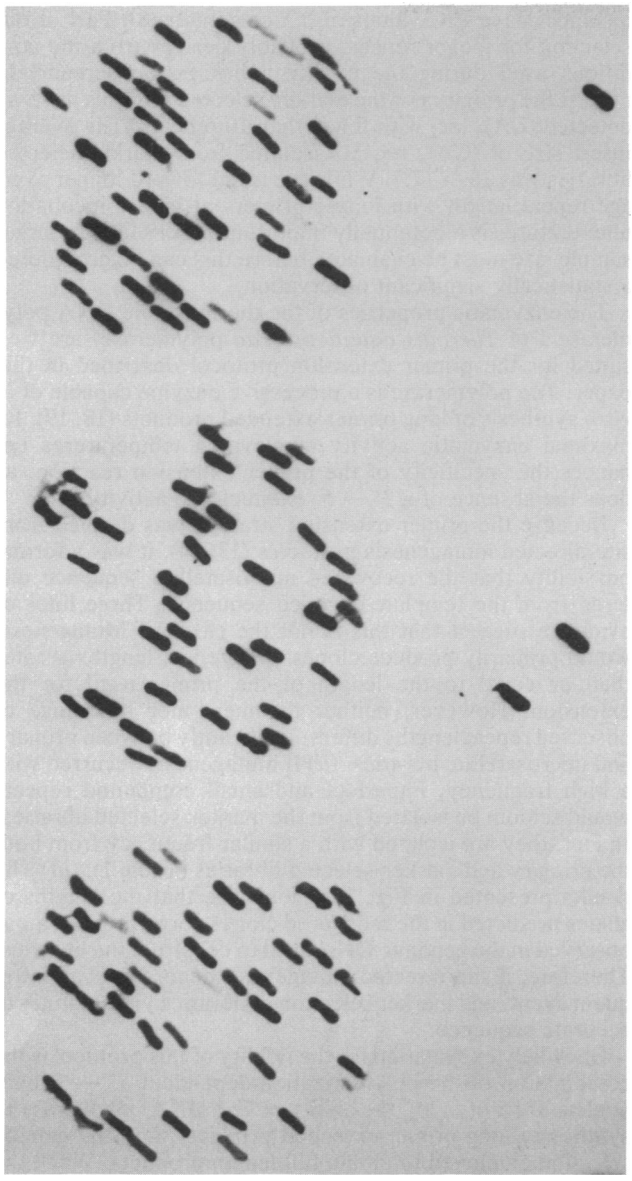

FIG. 1. Hybridization screening of random clones from primary and marker-selected libraries with a $(CA)_{15}$ oligonucleotide. Each of the three filters on the right contained 100 randomly chosen clones from a dog primary genomic library. Each of the three filters on the left contained 100 randomly chosen clones from a marker-selected library in which primer extension reactions were at 75°C with a $(CA)_{15}$ oligonucleotide. The filters were processed by incubation for 5 min on 3MM Whatman paper presoaked with 10% SDS, followed by denaturation (0.5 M NaOH/1.5 M NaCl), renaturation (0.5 M Tris, pH 8.0/1.5 M NaCl), and a rinse [2× standard saline citrate (SSC)], then air-dried and baked at 80°C for 60 min in a vacuum oven. Radiolabeled $(CA)_{15}$ oligonucleotide was prepared by incubating 100 pmol of oligonucleotide for 30 min at 37°C with 25 μCi of [γ-^{32}P]ATP (3000 Ci/mmol, Amersham; 1 Ci = 37 GBq) and 1 unit of T4 polynucleotide kinase in the buffer provided by the vendor (United States Biochemical), followed by incubation at 75°C for 10 min to destroy remaining kinase activity. Approximately 200 pmol of this probe was sufficient to screen six filters. Prehybridizations were carried out for 10 min at 65°C in 6× SSC/0.1% SDS. The probe was then added directly to the filters and hybridized for 45 min at 65°C. Hybridization reactions were cooled slowly to room temperature, and the filters were washed with 6× SSC/0.1% SDS at 55°C. Positive clones were identified following autoradiography at −70°C for 3–8 hr.

1). $(CA)_n$ repeats were counted and categorized by standard convention (8). The mean length value of repeats presented here is the arithmetic average of the longest continuous repeat sequence for each clone.

Table 1. Distribution of $(CA)_n$·$(GT)_n$ repeats in primary and marker-selected libraries

Primary library	Marker-selected libraries		
	n = 10	n = 15	n = 20
$(TG)_{12}$	$(CA)_{12}$	$(AC)_{10}$	$(AC)_{15}$
$(TG)_{13}$	$(CA)_{12}$	$(AC)_{11}$	$(AC)_{15}$
$(CA)_{14}$	$(CA)_{12}$	$(AC)_{15}$	$(AC)_{16}$
$(CA)_{14}$	$(TG)_{12}$	$(AC)_{16}$	$(CA)_{16}$
$(GT)_{15}$	$(CA)_{13}$	$(AC)_{16}$	$(AC)_{17}$
$(AC)_{15}$	$(AC)_{15}$	$(AC)_{16}$	$(CA)_{18}$
$(CA)_{16}$	$(TG)_{17}$	$(AC)_{17}$	$(CA)_{18}$
$(AC)_{16}$	$(GT)_{19}$	$(AC)_{18}$	$(AC)_{18}$
$(GT)_{16}$	$(GT)_{19}$	$(AC)_{19}$	$(CA)_{20}$
$(TG)_{17}$	$(AC)_{20}$	$(CA)_{20}$	$(AC)_{20}$
$(TG)_{19}$	$(AC)_{20}$	$(CA)_{21}$	$(AC)_{22}$
$(AC)_{20}$	$(AC)_{21}$	$(AC)_{22}$	$(CA)_{25}$
$(CT)_4A(GT)_{16}$	$(AC)_{22}$	$(AC)_3TG(AC)_{21}$	$(AC)_{18}AA(TC)_4$
$(TG)_{15}(TC)_5$	$(TTG)_3TT(GT)_{13}$	$(CA)_9(GA)_5$	$(AC)_{18}AT(CA)_3$
$(CT)_4(GT)_{16}$	$(AT)_{20}(CA)_{19}$	$(AT)_{12}(AC)_{13}$	$(AC)_{10}(AG)_{11}$
	Mean continuous repeat length		
15.6	16.4	16.3	17.7

Fifteen clones identified by colony hybridization were sequenced from the primary library and from each of the three marker-selected libraries. Primer extension reactions were done with $(CA)_n$ or $(TG)_n$ oligonucleotides in which n = 10, 15, or 20 as indicated. Mean repeat lengths for markers sequenced from each library are indicated. For compound and imperfect repeats, the longest continuous stretch of dinucleotides was used in calculating the mean. For probes with 15 or 20 repeats, only clones generated by $(CA)_n$ primers were sequenced. Single-stranded DNA was prepared for sequencing reactions by a slight modification of the method described by Vieira and Messing (15). Templates were analyzed by the dideoxynucleotide chain-termination method (17) using Sequenase DNA sequencing kits (United States Biochemical) and T7 primer (Stratagene).

The mean repeat length of a $(CA)_n$ microsatellite in the primary library was 15.6 with a range of 12–20 (Table 1). Analysis of repeat lengths from a *Sau*3A1 genomic DNA library with a larger average insert size produced similar results (data not shown). Similar analysis of marker-selected libraries constructed by using $(CA)_n$ and $(TG)_n$ 10-mer or 15-mer primers for synthesis of the complementary strand resulted in average repeat lengths of 16.4 and 16.3, respectively. The range of repeat lengths was 12–22 for libraries made with $(CA)_{10}$ and $(TG)_{10}$ primers and was 10–22 for libraries made with $(CA)_{15}$ and $(TG)_{15}$ primers. For marker-selected libraries constructed with a $(CA)_{20}$ primer, the mean repeat length was 17.7, with a range of 15–25.

PCR Amplification of $(CA)_n$ Repeats Identified in Marker-Selected Libraries. The strategy used for the construction of the marker-selected genomic DNA libraries is similar to a common method for site-directed mutagenesis (13, 14). Since the uracil-containing single-stranded template DNA was primed with a long oligonucleotide, it was formally possible that the microsatellite repeat observed in the recovered clones was mutated by the procedure, resulting in marker-selected libraries with artificially lengthened $(CA)_n$ repeats. This possibility was tested experimentally in the following way: PCR primers were designed from unique DNA flanking several of the repeats, and PCR products were then amplified directly from the genomic DNA of the single dog used to construct the primary libraries. The size of those products was compared with that of PCR products amplified from individual $(CA)_n$-containing clones identified in the screen of the marker-selected libraries (Fig. 2). For the four markers shown, as well as for one other tested, the size of the PCR products amplified from genomic DNA matched the size of

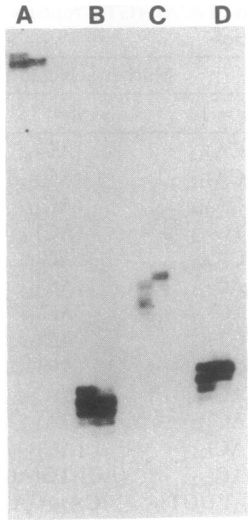

FIG. 2. Comparison of the lengths of PCR products prepared by amplification of genomic and plasmid DNA. For each of four markers (A–D) PCRs were done on genomic DNA prepared from the same dog used to construct the primary library (left lane in each case) and purified plasmid DNA from the appropriate clone from the marker-selected library (right lane). Markers: A, $(AC)_{15}$; B, $(GT)_{18}$; C, $(AC)_{23}$; D, $(AC)_{18}(AT)(AC)_3$. Markers A–C were isolated from a library made by primer extension with a $(CA)_{16}$ oligonucleotide; marker D was from a library prepared by primer extension with a $(CA)_{20}$ oligonucleotide. Sequences (5′ → 3′) of PCR primers were as follows: marker A, TATTAAATCCCAGTCACCACCC and AGGTCCCAGACCGAGTCC; marker B, ATCGCCTTAGTGCATGCAG and GTTTGGGTTTGGTAACATAGGC; marker C, AGCAACCCCTCCCATTTACT and TTGATCTGAATAGTCCTCTGCG; marker D, AATGGCAGGATTTTCTTTTGC and ATCTTTGGACGAATGGATAAGG.

the counterpart amplified from the microsatellite plasmid library clone to within 1–2 bp. This suggests that the method used to produce the microsatellite libraries has little, if any, mutagenic effect on $(CA)_n$ repeat length.

Several of the markers recovered from the microsatellite library have been tested on populations of unrelated hybrid dogs in order to determine the utility of this set of markers for further study. In most cases multiple alleles have been detected. A more comprehensive survey of these markers and the distribution of polymorphisms will be presented elsewhere.

DISCUSSION

We have described an efficient protocol for the rapid production of genomic DNA libraries highly enriched for microsatellite repeats. In this protocol, a strong genetic selection was employed to identify a desired class of clones from a complex mammalian library. The resulting marker-selected small genomic DNA libraries are 50-fold enriched for the targeted products. Since the primary library was constructed in phagemid vectors allowing the rapid isolation of single-stranded DNA, sequence of the small inserts suitable for the design of polymorphic STSs was readily obtained. Experiments comparing the size of PCR products amplified from genomic DNA and plasmids from the marker-selected libraries suggested that marker selection of libraries was largely nonmutagenic.

Colony hybridization of both the primary and marker-selected libraries with a $(CA)_{15}$ oligonucleotide identified relatively long repeats. The available published data for human DNA suggest that the length of the repeat may be predictive of the potential PIC value of a randomly selected marker. $(CA)_n$ repeats that contain 16 or more repeats, for instance, have PIC values of 0.5 or greater (8). Part of our selection for longer repeats probably derives from the conditions used during the colony hybridization screens. In neither the primary nor the marker-selected libraries have we detected $(CA)_n$ loci with fewer than 10 repeats. The average insert size of $(CA)_n$ repeats isolated from marker-selected libraries was 16.3–17.7. While the trend toward longer average repeat length with longer primers at higher incubation temperatures is a potentially important observation, a larger sample size must be evaluated before this can be considered a statistically significant observation.

The enzymatic properties of the thermostable DNA polymerase I of *Thermus aquaticus* (*Taq* polymerase) are well suited for the primer extension protocol described in this paper. *Taq* polymerase is a processive enzyme capable of *in vitro* synthesis of long primer-extended products (18, 19). Its maximal enzymatic activity at elevated temperatures enhances the specificity of the primer extension reaction, as does the absence of a 3′ → 5′ exonuclease activity.

Because the primer extension strategy was derived from site-directed mutagenesis protocols (13, 14), it was a formal possibility that the recovered microsatellite sequence differed from the template-encoded sequence. Three lines of evidence suggest that this is not the case. (*i*) Mutagenesis would primarily produce clones with repeat lengths greater than or equal to the length of the primer used for the extension. However, neither the mean nor the range of observed repeat lengths differs significantly between primary and microsatellite libraries. (*ii*) If mutagenesis occurred with a high frequency, imperfect and small compound repeats would seldom be isolated from the marker-selected libraries. In fact, they are isolated with a similar frequency from both the primary and marker-selected libraries (Table 1). (*iii*) The results presented in Fig. 2 demonstrate that the lengths of alleles predicted in the recovered clones correspond to those observed in the genomic DNA used to construct the libraries. Therefore, if site-directed mutagenesis occurs, it is an infrequent event, and marker selection of libraries yields clones of accurate sequence.

One likely explanation for the fidelity of this protocol is the combination of the polymerization-dependent 5′ → 3′ exonuclease activity (20), the ability of *Taq* DNA polymerase to synthesize long primer-extended products, and the genetic selection employed to obtain full-length products. When the *Taq* DNA polymerase completes a round of synthesis using the circular single-stranded DNA template, the enzyme encounters the 5′ end of the primer–template complex. At this point, particularly if there is incorrect base pairing between the primer and template, the 5′ → 3′ exonuclease activity of the enzyme may remove the primer. The template-encoded microsatellite will then serve as a template for faithful replication. The resultant nick is sealed *in vitro* by addition of T4 DNA ligase prior to transformation of the products into bacterial cells. The ligation step should help select for molecules in which gap-repair has been completed. It should be noted that both duplex primer–template complexes as well as duplex complexes with 5′ single-stranded DNA extensions can be substrates for the 5′ → 3′ exonuclease activity of *Taq* polymerase.

Under ideal circumstances, the marker-selected libraries would include only clones containing the targeted microsatellite, yet empirically half of the clones do not contain the targeted repeat. Control experiments in which primer extension reactions were performed in the absence of either primer or *Taq* polymerase suggest that a portion of the background arises from random priming by contaminating *E. coli* chromosomal DNA or RNA. Alternative methods for preparation of single-stranded DNA may minimize this problem.

A potential drawback to this procedure is that the amplification step may skew the representation of loci in the

marker-selected library. For example, the pooling of the primary genomic library followed by infection with helper M13 phage may lead to the loss of some clones by differential replication. The use of small-insert libraries and attention to the details of protocols for cell and phage growth should help to limit this problem. Another drawback is that the commonly available *dut ung* strains, including the one used in this study, contain restriction–modification systems.

Other methods, based on affinity chromatography, have recently been described as a means for collecting populations of markers enriched for $(CA)_n$ repeats (21). To date, however, these approaches have been used to generate limited libraries in which only 10% of the members have $(CA)_n$ repeats. In addition, these libraries identify repeats in the range of $(CA)_3$ to $(CA)_{17}$ and, thus, on average, are much smaller than the repeats identified in this protocol.

The primer extension protocols as outlined in this paper used either $(CA)_n$ or $(TG)_n$ primers. In principle, all classes of simple sequence repeats, including trimeric and tetrameric tandem repeats (22), are potential targets. Since 25,000 clones containing $(CA)_n$ repeats could be recovered per microgram of input DNA from this procedure, it is likely that classes of sequence repeats present 100–1000 times less frequently than $(CA)_n$ repeats could also be recovered.

This strategy was originally conceived as a rapid method for the generation of a large number of random markers, and as a consequence total genomic DNA was used as the source for libraries. Since the construction of high-resolution genetic maps will require the production of large numbers of chromosome-specific or region-specific markers, libraries should be constructed with flow-sorted chromosomes and pooled, chromosome-specific large-insert libraries as starting materials. The use of this protocol, combined with high-throughput sequencing strategies, will lead to the rapid accumulation of genetically useful STSs and fuel the development of robust genetic maps.

We thank Penny Mapa for excellent technical assistance and George Sprague, David Cox, Richard Myers, and Chancellor Helmut Kohl for their contributions to this work. We also acknowledge George Sprague and George Church for critical reading of this manuscript. This work was supported by a grant from the Lucille P. Markey Charitable Trust (J. Thorner and D. Koshland, principal investigators) and by a National Institute of Environmental Health Sciences Mutagenesis Center Grant to J.R. (ESO1896). G.D. was a Fellow of the Lucille P. Markey Charitable Trust at the time this work was initiated.

1. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32,** 314–331.
2. Litt, M. & Luty, J. A. (1989) *Am. J. Hum. Genet.* **44,** 397–401.
3. Weber, J. A. & May, P. E. (1989) *Am. J. Hum. Genet.* **44,** 388–396.
4. Miesfeld, R., Krystal, M. & Arnheim, N. (1981) *Nucleic Acids Res.* **9,** 5931–5947.
5. Hamada, H. & Kakunaga, T. (1982) *Nature (London)* **298,** 396–398.
6. Hamada, H., Petrino, M. G. & Kakunaga, T. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6465–6469.
7. Tautz, D. & Renz, M. (1984) *Nucleic Acids Res.* **12,** 4127–4138.
8. Weber, J. (1990) *Genomics* **7,** 524–530.
9. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245,** 1434–1435.
10. Weber, J. (1990) in *Genome Analysis Volume 1: Genetic and Physical Mapping* (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 159–181.
11. Green, E. D., Mohr, R. M., Idol, J. R., Jones, M., Buckinham, J. M., Deaven, L. L., Moyzis, R. K. & Olson, M. V. (1991) *Genomics* **11,** 548–564.
12. Ausubel, F. M., Brent, R., Kinston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1987) *Current Protocols in Molecular Biology* (Wiley, New York).
13. Kunkel, T. A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 488–492.
14. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. (1987) *Methods Enzymol.* **154,** 367–383.
15. Vieira, J. & Messing, J. (1987) *Methods Enzymol.* **153,** 3–11.
16. Lindahl, T. (1982) *Annu. Rev. Biochem.* **51,** 61–80.
17. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
18. Lawyer, F. C., Stoffel, S., Saiki, R. K., Myambo, K., Drummond, R. & Gelfand, D. H. (1989) *J. Biol. Chem.* **264,** 6427–6437.
19. Innis, M. A., Myambo, K. B., Gelfand, D. H. & Brow, M. A. D. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9436–9440.
20. Holland, P., Abramson, R. D., Watson, R. & Gelfand, D. H. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 7276–7280.
21. Brenig, B. & Brem, G. (1991) *Nucleic Acids Res.* **19,** 5441.
22. Edwards, A., Civitello, A., Hammond, H. A. & Caskey, C. T. (1991) *Am. J. Hum. Genet.* **49,** 746–756.